Soft-biometrics: Unconstrained Authentication in a Surveillance Environment

Simon Denman, Clinton Fookes, Alina Bialkowski, Sridha Sridharan
Image and Video Laboratory
Queensland University of Technology
Brisbane, Australia
s.denman@qut.edu.au, c.fookes@qut.edu.au, a.bialkowski@student.qut.edu.au, s.sridharan@qut.edu.au

***Abstract*—Soft biometrics are characteristics that can be used to describe, but not uniquely identify an individual. These include traits such as height, weight, gender, hair, skin and clothing colour. Unlike traditional biometrics (i.e. face, voice) which require cooperation from the subject, soft biometrics can be acquired by surveillance cameras at range without any user cooperation. Whilst these traits cannot provide robust authentication, they can be used to provide coarse authentication or identification at long range, locate a subject who has been previously seen or who matches a description, as well as aid in object tracking. In this paper we propose three part (head, torso, legs) height and colour soft biometric models, and demonstrate their verification performance on a subset of the PETS 2006 [1] database. We show that these models, whilst not as accurate as traditional biometrics, can still achieve acceptable rates of accuracy in situations where traditional biometrics cannot be applied.**

***Keywords*-Soft Biometric, Appearance Model, Surveillance**

I. Introduction

Current video surveillance systems do little more than acquire footage for review by human operators. In a typical scenario one or two operators would be responsible for tens or even hundreds of different cameras. Due to the massive number of cameras deployed, it is no longer possible or efficient to have human operators continuously and accurately monitoring the multitude of video feeds at all times. Consequently, the likelihood of important events being detected as they happen, or people of interest being located is extremely low. Several recently developed intelligent surveillance systems and commercial products now make it possible to track individuals. These systems are quite robust in scenes with little clutter and small amounts of occlusion. However, they are inadequate for tracking large numbers of people in heavily crowded environments, or locating and tracking a person of interest through a crowded scene.

Soft biometrics are characteristics that can be used to describe, but not uniquely identify an individual. Traits such as height, weight, gender, hair, skin and clothing colour are examples of soft biometrics. Unlike traditional biometrics (i.e. face, voice) which require cooperation from the subject, soft biometrics can be acquired by surveillance cameras at range without any user cooperation. Whilst these traits cannot provide robust authentication, they can be used to provide coarse authentication or identification at long range, locate a subject who has been previously seen or who matches a description, as well as aid in object tracking.

For soft biometrics to be of use in a real world surveillance environment, they must be view invariant, and robust to illumination changes [2]. Jain et al [3] demonstrated that soft biometrics such as height, gender and ethnicity can be used to improve performance of a traditional biometric system. Ran et al [2] proposed a gait signature, consisting of several soft biometrics based on gait features. Stride length, height and gender could all be extracted from a video sequence and it was shown that these features are effective for limited recognition.

Appearance modeling techniques used in object tracking systems can also be used as soft biometrics. Appearance models are typically designed to be view and illumination invariant so that they may be used to aid in tracking handover between different camera views, and to aid in tracking during or after occlusions. Haritaoglu et al [4] proposed a method where data pertaining to the average texture and silhouette of the subject is recorded over a period of time as the object is tracked. This model can be used to determine the identity of a person who has just ceased to be occluded, or can be used to re-detect a person if they had been lost for several frames due to occlusion, or had left and re-entered the scene.

Whilst the approach proposed in [4] is effective for a single view, texture information is not suitable for transferring from one view to another. Hu et al [5] extracted three histograms from each person, one each for the head, torso and legs, to not only allow for matching based on colour, but also on distribution of colour. Chien et al [6] proposed a colour model (Human Colour Structure Descriptor - HCSD) that aims to capture the distribution of colours in a human body. Three colours are used to represent the colour of the body, legs and shoes, and positions are defined to describe the position of body and legs relative to the shoes.

Nakajima et al [7] and Hahnel et al [8] each proposed techniques to model and recognise people based on their whole body. Nakajima et al [7] extracts normalised colour histograms and local shape features for detected people and trains SVM classifiers for each person and pose and the approach is shown to be accurate on a limited data set. Hahnel et al [8] extends on [7] by applying additional colour, shape and texture features.

In this paper we propose three part (head, torso, legs) height and colour soft biometric models, and demonstrate their verification performance on a subset of the PETS 2006 [1] database. We show that these models, whilst not as accurate as traditional biometrics, can still achieve acceptable rates of accuracy in situations where traditional biometrics cannot be applied. The remainder of this paper is structured as follows: Section II describes the soft biometric models proposed; Section III outlines the test database and testing procedure; Section IV shows test results and Section V concludes the paper and discusses possible future work.

II. Soft Biometric Models

Soft biometrics are features that can be easily extracted from a distance. Ideally, for use in a surveillance environment any features should also be view invariant. In this paper, a model that consists of a size and a colour component to model a person is proposed. The person is segmented in head, torso and legs and each section is treated separately. The segmentation process is described in Section II-A, the size models in Section II-B and the colour models in Section II-C.

The proposed model assumes that people appear vertically in the image, and that there is no significant optical distortion resulting in an abnormal appearance. It is also assumed that the cameras used are calibrated using a technique that allows image coordinates to be translated into a common world coordinate scheme (i.e. Tsai [9]).

A. Person Segmentation

An adaptive background segmentation routine [10] is used to separate any moving objects from the background, and detect the person. Figure 1 shows a cropped input image and the corresponding motion image.

(a) Colour (b) Motion

Figure 1. Using Background Segmentation to Separate a Person from the Scene

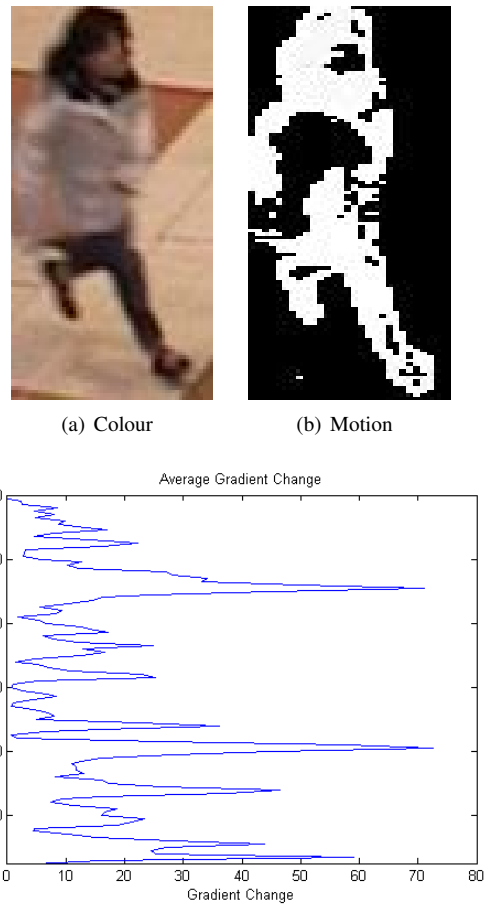
People are segmented into head, torso and leg regions. Segmentation is performed by analysing the average vertical gradient for each row, and locating maxima. Typically, when analysing the colour image of a person, significant colour changes are observed at the neck (top of a shirt) and the waist (boundary between a shirt and pants). Depending on the appearance of the person (i.e. complex pattens on clothing, single garment worn on the whole body etc.) this approach may fail, however other potential approaches that use shape are more likely to be view dependant.

Average gradient for each row is computed as follows,

$$v_{grad}(j) = \frac{\sum_{i=1}^{i=N-1} |I(i,j) - I(i,j-1)| \times M(i,j)}{\sum_{i=0}^{i=N-1} M(i,j)}, \quad (1)$$

where v_{grad} is a vector of the average gradient, j is the row being operated on, N is the width of inputs images, $I(i,j)$ is the colour image and $M(i,j)$ is the binary motion image (0 for no motion, 1 for motion).

From this vector of average gradients, the person can be segmented into head, torso and legs. Figure 2 shows an example of the average horizontal gradient over the height of a person.



(a) Colour (b) Motion

(c) Gradient

Figure 2. Average Horizontal Gradient for a Person

It is assumed that the person is orientated vertically in the image, that there are no significant errors in the motion segmentation, and that there is no significant distortion such that the person appears normally proportioned.

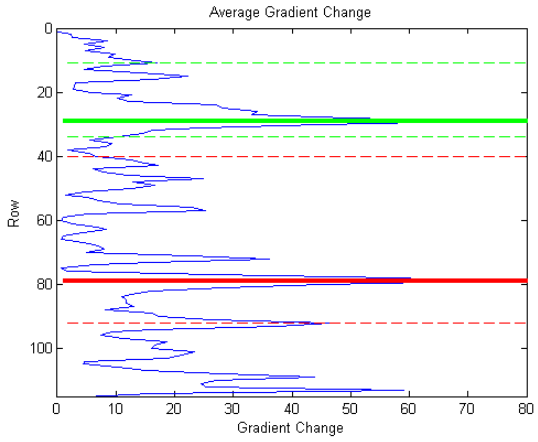
Given the aforementioned assumptions, the maxima for the neck should be at between 10% and 30% of the person's height. The neck is located at,

$$P_{neck} = argmax_{i=0.1\times H}^{0.3\times H}(v_{grad}(i)), \quad (2)$$

where P_{neck} is the position of the row denoting the neck in pixels (with 0 being the top of the person), and H is the height of the person in pixels. For the waist, it is assumed that it should lie at between 35% and 70% of the person's height. Given this, the waist is located at,

$$P_{waist} = argmax_{i=0.35\times H}^{0.7\times H}(v_{grad}(i)), \quad (3)$$

where P_{waist} is the position of the row denoting the waist in pixels (with 0 being the top of the person). Figure 3 shows the detected boundaries for a person.



(a) Colour

(b) Gradient

Figure 3. Detection of Neck (Green) and Waist (Red) - The dotted lines indicate the search bounds, the solid lines the detected boundary.

If the neck or waist cannot be found (i.e. failure of motion segmentation or person detection), the images are discarded and no further computation takes place.

B. Size

The height of the person is used as a simple descriptor. The height is view invariant, whilst other dimensions (width and length) are dependent on the camera angle as well as the persons pose (i.e. as a person walks their width changes as their legs move). Heights are stored for head, torso and legs.

To determine the height of the person, the head and feet must be located in the image. The top of the head is located by searching the motion image for the person to determine the highest point on the top contour, x_h, y_h. The feet position, y_f, is determined by finding the average height of the bottom contour for all pixels on the bottom contour that are within with bottom 20% of the image,

$$B_{contour}(i) = maximum\ j\ for\ which\ M(i,j) > 0, \quad (4)$$

$$y_f = \frac{\sum_{i=0}^{X-1} B_{contour}(i)\ where B_{contour}(i) > Y \times \tau}{B_{contour}(i) > Y \times \tau} \quad (5)$$

where $B_{contour}$ is the bottom contour of the motion image, M, X and Y are the width and height of M respectively, and τ is a height threshold used to determine which parts of the contour lie on the ground plane (set to 0.8 as we are using the parts of the contour that lie within 20% of the bottom of the image). It is assumed that for the motion image (M) it is zero indexed and the top left corner is at the coordinate $(0, 0)$. Figure 4 shows an example of the located head and feet points for a person.

(a) Colour Image (b) Detected Head and Feet

Figure 4. Detecting the Head and Feet

This approach for locating the feet aims to find the average height of the two feet. By restricting the average to the lower 20% of the image, is prevents poses such as those where the person is walking and has their legs far apart from distorting the results. Figure 5 shows an example of this. It can be seen that when the whole of the bottom contour is considered, it can distort results under certain conditions.

The x-coordinate of the feet is set to that of head, x_h (for a person standing vertically, the feet should be directly below the head). Neck and torso boundaries are determined

(a) Colour Image (b) $\tau = 0$ (c) $\tau = 0.8$

Figure 5. Effect of τ - In (b), the height of the feet in the image is incorrect.

as described in Section II-A, and coordinates for these boundaries are set to x_h, P_{neck} and x_h, P_{waist} respectively. Using camera calibration, the image coordinates can be transferred into a real world coordinate scheme, and head, torso and leg heights can be determined,

$$H_{head} = z^w_{head} - z^w_{neck}, \tag{6}$$
$$H_{torso} = z^w_{neck} - z^w_{waist}, \tag{7}$$
$$H_{legs} = z^w_{waist} - z^w_{feet}, \tag{8}$$

where H_{head}, H_{torso} and H_{legs} are the head, torso and legs heights in world coordinates, and z^w_{head}, z^w_{neck}, z^w_{waist}, z^w_{feet} are the real world z-coordinates (height off the ground plane) of the head, neck, waist and feet. z^w_{feet} is always set to 0 (i.e. the person's feet are on the ground).

Heights are progressively updated over multiple observations,

$$H'_{head}(t) = \frac{L-1}{L} \times H'_{head}(t-1) + \frac{H_{head}(t)}{L}, \tag{9}$$
$$H'_{torso}(t) = \frac{L-1}{L} \times H'_{torso}(t-1) + \frac{H_{torso}(t)}{L}, \tag{10}$$
$$H'_{legs}(t) = \frac{L-1}{L} \times H'_{legs}(t-1) + \frac{H_{legs}(t)}{L}, \tag{11}$$

where $H'_{head}(t)$, $H'_{torso}(t)$ and $H'_{legs}(t)$ are the average head, torso and leg heights for the model at time t; $H_{head}(t)$, $H_{torso}(t)$ and $H_{legs}(t)$ are the heights for the image at the current time step computed as described in Equations 6 to 8, and L is the learning rate. L is defined as,

$$L = \frac{1}{T}; for\ T < W, \tag{12}$$
$$L = \frac{1}{W}; for\ W >= T, \tag{13}$$

where W is the number of frames used in the model, and T is the number of updates performed on the model. This ensures that the image that the model is initialised with does not dominate the model for a significant number of frames. Instead, new information is incorporated quickly when the model is new to provide a better representation of the object being modeled sooner.

An error measure is kept for the heights,

$$F^e_{head}(t) = |H'_{head}(t) - H_{head}(t)|, \tag{14}$$
$$F^e_{torso}(t) = |H'_{torso}(t) - H_{torso}(t)|, \tag{15}$$
$$F^e_{legs}(t) = \left|H'_{legs}(t) - H_{legs}(t)\right|, \tag{16}$$

where $F^e_{head}(t)$, $F^e_{torso}(t)$ and $F^e_{legs}(t)$ are the frame errors for the head, torso and leg heights. The errors are updated over time using equations 12 and 13, to generate the average errors, E_{head}, E_{torso} and E_{legs}, for the head, torso and legs respectively. The cumulative error is used as an approximation to the standard deviation (it is assumed that the observations over time form a Gaussian distribution) of the error, as it is not practical to re-compute the standard deviation each frame, and not ideal to assume a fixed standard deviation. Given that the standard deviation for a sample set is defined as,

$$\sigma = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\mu - s_n)^2}, \tag{17}$$

and in the proposed model, for each measure there is one observation at each time step ($N = 1$), so the standard deviation at a given time step is,

$$\sigma = \sqrt{(\mu - s)^2} = |H'(t) - H(t)|, \tag{18}$$

which is the proposed error measure.

When comparing two size models, the mean heights and approximated standard deviations are used to determine the probability of a match. The probability for head, torso and legs heights are defined as,

$$P_{head}(i,j) = \Phi_{0,E_{head}(i)}(|H'_{head}(i) - H'_{head}(j)|), \tag{19}$$
$$P_{torso}(i,j) = \Phi_{0,E_{torso}(i)}(|H'_{torso}(i) - H'_{torso}(j)|), \tag{20}$$
$$P_{legs}(i,j) = \Phi_{0,E_{legs}(i)}(\left|H'_{legs}(i) - H'_{legs}(j)\right|), \tag{21}$$

where $P_{head}(i,j)$ is match between the head component two models, i and j, $E_{head}(i)$ is the approximated standard deviation of the head height for model i, $H'_{head}(i)$ is the mean head height for model i, and$\Phi_{\mu,\sigma}$ is the cumulative density function for the Gaussian distribution. The average of these scores,

$$P_{height}(i,j) = \frac{P_{head}(i,j) + P_{torso}(i,j) + P_{legs}(i,j)}{3}, \tag{22}$$

is taken as the match between models i and j.

C. Colour

Colour histograms are computed for the head, torso and leg sections (C_{head}, C_{torso} and C_{legs} respectively). The colour and motion image are used to generate histograms, such that only pixels that are in motion (i.e. part of the person) are included in the histogram.

A moving average of the histogram is calculated such that,

$$C'(t) = \frac{L-1}{L} \times C'(t-1) + \frac{C(t)}{L}, \tag{23}$$

where $C'(t)$ is the value of the average histogram at time t, $C(t)$ is the histogram computed for the frame at time t, and L is the learning rate. L is set as shown in Equations 12 and 13.

Histograms are compared using the Bhattacharya coefficient,

$$B(C^i, C^j) = \sqrt{\sum_{1}^{N} \sqrt{C^i(n) \times C^j(n)}}, \tag{24}$$

where $B(C^i, C^j))$ is the Bhattacharya coefficient that results for the comparison of the histograms C^i and C^j, $C^i(n))$ is the nth bin for the histogram C^i, and N is the total number of bins in the histogram. The histogram comparison is performed using histograms with their bin weights normalised such that they sum to 1,

$$\sum_{1}^{N} C^i(n) = 1. \tag{25}$$

This is done to ensure size invariance. The comparison will return 1 for a perfect match, and 0 for no match.

When comparing colour models for two people, the similarity score is taken as the average of the three histogram comparisons,

$$P_{colour}(i,j) = \frac{1}{3} \times (B(C^i_{Head}, C^j_{Head}) + B(C^i_{Torso}, C^j_{Torso}) + B(C^i_{Legs}, C^j_{Legs})), \tag{26}$$

where $C(i,j)$ is the similarity score between models i and j.

D. Combining Soft Biometric Models

As both the size and colour soft biometrics measure different features, it is logical to combine them when modeling people. A simple weighted sum fusion approach is proposed,

$$P(i,j) = \alpha P_{colour}(i,j) + (1-\alpha) P_{size}(i,j), \tag{27}$$

where α is a weight used to combine the models.

III. Test Database

A test database is formed using a portion of the PETS 2006 [1] database. PETS 2006 [1] is a four camera database captured at a train station for detecting abandoned objects. Calibration information is provided for the four cameras. Of the four cameras, only cameras 3 and 4 are used in the proposed test database. These cameras have a significant overlap and are both positioned high above the ground to limit occlusions between people. Camera 1 is mounted very low to the ground meaning occlusions are a significant problem, and Camera 2 is mounted much further from the area of overlap. An example of the four camera feeds can be seen in Figure 6.

The database consists of sets of 10 consecutive frames where a person is clearly visible (i.e. is not obscured by other people) in both cameras 3 and 4. Two sets of 10 frames are taken for each person, such that for each person there are four sets of 10 frames. These sets are spaced as widely as possible apart, however the temporal separation between the sets depends on how long the person is present in both cameras. Depending on the trajectory the person has taken, this varies between 0 and several hundred frames. Data is taken for 25 different people. The size of the database is limited to 25 subjects due to the nature of the PETS 2006 [1] data. To provide a fair evaluation of the proposed biometrics, only subjects who are unobstructed and entirely within the image bounds in both views simultaneously are included in the database. As a result, many of the people in PETS 2006 are unsuitable. The database structure is shown in Table I.

Subject	Session	Camera	Number of Frames
01	1	3	10
	1	4	10
	2	3	10
	2	4	10
02	1	3	10
	1	4	10
	2	3	10
	2	4	10
.....			
25	1	3	10
	1	4	10
	2	3	10
	2	4	10

Table I
TEST DATABASE STRUCTURE

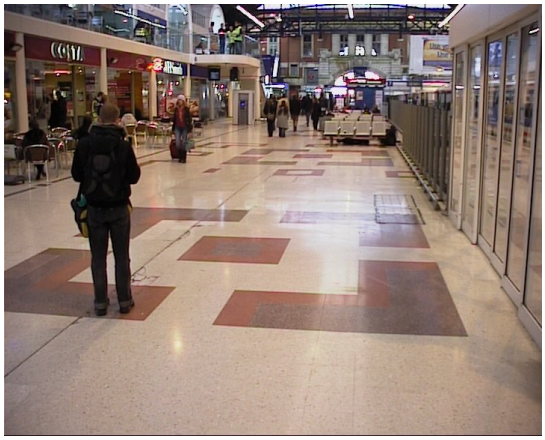
Each set of frames is hand annotated with the bounding box of the person. Whilst it is possible to automatically extract person bounds using person detection and object tracking techniques, hand annotation ensures repeatable tests, and that errors observed are not the result of poor detection and tracking. Motion detection is performed however as it is not feasible to manually segment each image. Figure 7 shows samples of the people in the database. Cropped images (cropped according to the annotation) of subjects 1 to 5 taken from the two camera views are shown. It can be seen that many of the subjects within the database are similarly dressed in dark clothing. This is typical of the people observed within the PETS 2006 database.

A. Test Configuration

The proposed soft biometric models are tested in two different situations:

1) Models are built from a single camera view.
2) Models are built from two camera views.

To test models built from a single view, the database test configuration shown in Table II is used.



(a) Camera 1

(b) Camera 2

(c) Camera 3

(d) Camera 4

Figure 6. PETS 2006 Camera Views - Simultaneous images from the four camera views, the person with the red luggage can be seen in each camera.

Training	Testing		
S1-C3	S1-C4	S2-C3	S2-C4
S1-C4	S1-C3	S2-C3	S2-C4
S2-C3	S1-C3	S1-C4	S2-C4
S2-C4	S1-C4	S1-C4	S2-C3

Table II
TEST CONFIGURATION FOR TESTING MODELS BUILT USING A SINGLE CAMERA VIEW

Training	Testing		
S1	S2	S3	S4
S2	S1	S3	S4
S3	S1	S2	S4
S4	S1	S2	S3

Table III
TEST CONFIGURATION FOR TESTING MODELS BUILT USING TWO CAMERA VIEWS

To test models built from two cameras, the four data sets for each subject are reconfigured into four sessions,

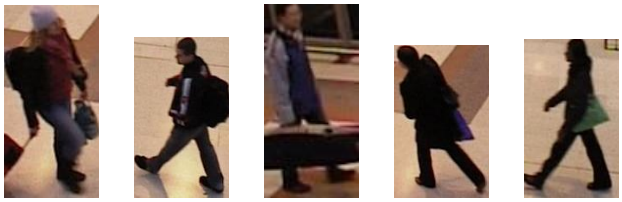
- X-S1 = [X-S1-C3-F1-5, X-S1-C4-F1-5],
- X-S2 = [X-S1-C3-F6-10, X-S1-C4-F6-10],
- X-S3 = [X-S2-C3-F1-5, X-S2-C4-F1-5],
- X-S4 = [X-S2-C3-F6-10, X-S2-C4-F6-10],

where X-S1-C3-F1-5 is the first session (S1), camera three (C3), frames 1 to 5 (F1-5), for subject X. Using these four sessions for each subject, the test configuration shown in Table III is used.

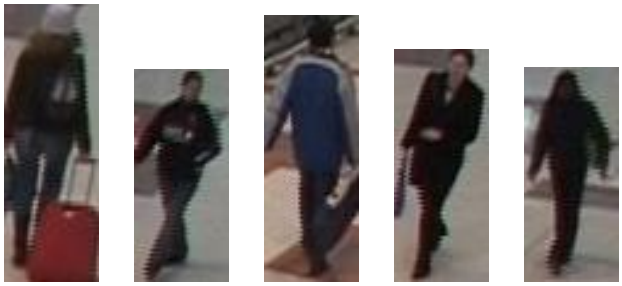
IV. RESULTS

Two sets of tests are performed as described in Section III-A. For each test, the individual colour and size biometrics as well the combined biometrics (see Sections II-C, II-B and II-D respectively) are evaluated. For the combined biometrics, $\alpha = 0.5$.

Figure 8 shows a DET plot for models trained on a single view (test configuration one, see Section III-A). Using the colour model alone performs best (equal error rate, 26.7%), although the combination of the colour and size model (equal error rate, 29.8%) performs better at some operating points. The size model alone (equal error rate, 39.6%) clearly



(a) Subject 1, Camera 3 (b) Subject 2, Camera 3 (c) Subject 3, Camera 3 (d) Subject 4, Camera 3 (e) Subject 5, Camera 3



(f) Subject 1, Camera 4 (g) Subject 2, Camera 4 (h) Subject 3, Camera 4 (i) Subject 4, Camera 4 (j) Subject 5, Camera 4

Figure 7. A Sample of the Test Database - Cropped images for the first five subjects taken from the two cameras are shown.

performs the worst of the three.

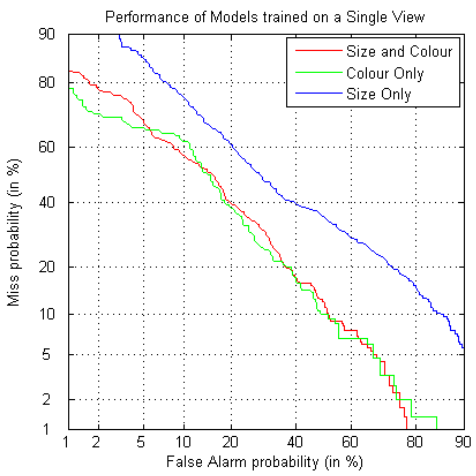


Figure 8. Verification Results for Models Trained on a Single View

This poor performance can in part be attributed to the difference between the two views, and poor performance when matching models trained on different cameras. Cameras three and four are positioned such that camera three typically views the subject from side on (or partially side on) while camera four views that subject from either the front or back (depending on which direction they are walking). The subjects also appear at very different sizes in the images (in camera three subjects are approximately 180-250 pixels tall, in camera four they are approximately 80-120 pixels tall). The use of calibrated cameras allows these measurements to be translated to a common coordinate scheme. However, the smaller size of the objects in camera four results in less accurate size models for subjects in this view. An error of a few pixels in localisation of the head and feet results in a larger error than it would in camera 3. When testing and training models are restricted so that comparisons are only made between the same view (i.e. models trained from camera 3 are only compared to models trained from camera 3) equal error rates drop to 13.1%, 9.9% and 28% for combined colour and size, colour only and size only respectively.

The poor performance of the size model is in part due to errors in feet localisation, and the similarity of heights observed in the data set. A localisation error of a few of pixels potentially results in several centimeters difference in computed height (this is particularly the case for camera 4 where people appear much smaller). Due to the simple nature of the fusion of the two soft biometric modalities, the combination of the two results in a performance drop when compared to the colour alone.

Figure 9 shows a DET plot for models trained on multiple views (test configuration two, see Section III-A). Using the colour model only once again performs best (equal error rate, 6.1%). The colour only approach significantly outperforms combination of the colour and size model (equal error rate, 14.7%), and the size model alone (equal error rate, 22.7%) is the worst performing.

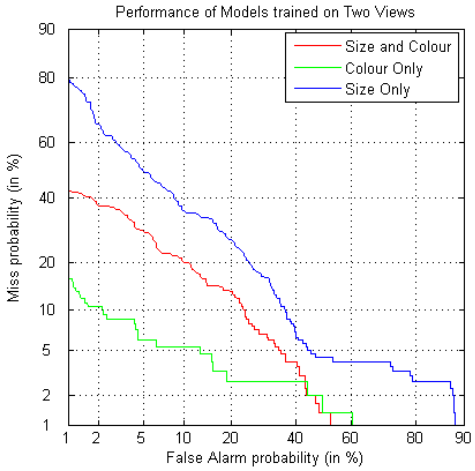
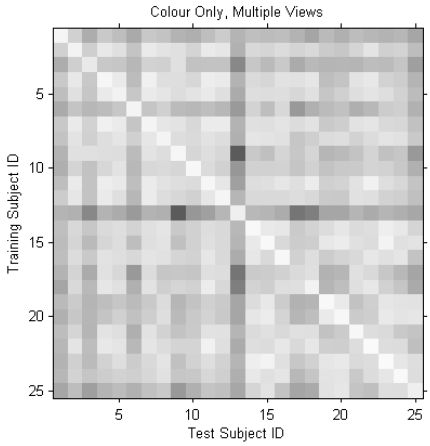


Figure 9. Verification Results for Models Trained on Multiple Views

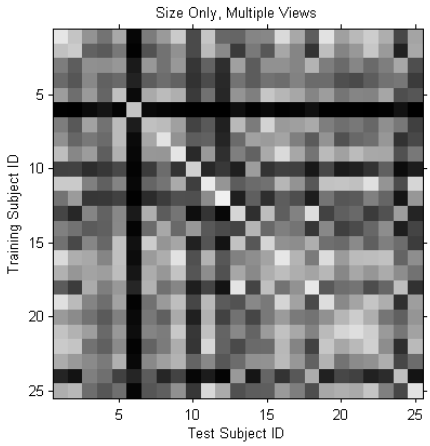
As can be seen in Figures 8 and 9, using multiple views for training models results is a significant increase in performance. Difference in colour distribution can be observed between the different cameras (i.e. a backpack may be more visible from one camera than another), and so building models using both views results in distributions that better describe individual people. The size model still

suffers from any segmentation errors, though does improve significantly by using an additional camera. Once again the simple fusion scheme is unable to effectively combine the two modalities.

Figure 10 shows confusion matrices for the models trained on two camera views (test configuration two, see Section III-A).



(a)



(b)

Figure 10. Confusion Matrices for Models Trained on Multiple Views

In Figure 10, it can be clearly seen that the colour modality is better able to discriminate between the subjects than the size modality. This is to be expected given the superior performance achieved by the colour modality. However, the size modality is in many cases able to either correctly identify the subject, or rank them highly. In some cases the size modality is able to better identify a subject than the colour modality (i.e. subject 6). Although the fusion approach used in this paper does not aid performance, it is expected that a more intelligent fusion scheme will be able to make better use of the two modalities and result in an increase in performance over the colour modality alone.

V. Conclusion

In this paper, we have proposed and evaluated a simple colour and size based soft biometric model for recognising people in a surveillance environment. Using a small database, we have shown that an equal error rate of 6.1% can be achieved for a recognition task when models are trained from multiple view points. Future work will focus on improving the proposed colour and size models as well as investigating additional modalities, and comparing the proposed biometrics with others. A new, larger, database will also be acquired, consisting of up to six cameras views of each subject to facilitate further research. The developed techniques will be evaluated on verification and identification tasks, and will also be used to aid in tracking objects in heavily crowded environments.

References

[1] J. M. Ferryman, Ed., *Proceedings of the Ninth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, June 2006.

[2] Y. Ran, G. Rosenbush, and Q. Zheng, "Computational approaches for real-time extraction of soft biometrics," in *IEEE Int. Conf. On Pattern Recognition*, 2008, pp. 1–4.

[3] A. K. Jain, S. C. Dass, and K. Nandakumar, "Soft biometric traits for personal recognition systems," in *International Conference on Biometric Authentication*, Hong Kong, 2008, pp. 731–738.

[4] I. Haritaoglu, D. Harwood, and L. Davis, "W4: real-time surveillance of people and their activities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 809 – 830, 2000.

[5] M. Hu, W. Hu, and T. Tan, "Tracking people through occlusions," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, 2004, pp. 724– 727.

[6] S.-Y. Chien, W.-K. Chan, D.-C. Cherng, and J.-Y. Chang, "Human object tracking algorithm with human color structure descriptor for video surveillance systems," in *Multimedia and Expo, 2006 IEEE International Conference on*, 2006, pp. 2097–2100.

[7] C. Nakajima, M. Pontil, B. Heisele, and T. Poggio, "Full-body person recognition system," *Pattern recognition*, vol. 36, no. 9, pp. 1997–2006, 2003.

[8] M. Hahnel, D. Klunder, and K. Kraiss, "Color and texture features for person recognition," in *IEEE International Joint Conference on Neural Networks*, Budapest, Hungary, 2004, p. 652.

[9] R. Y. Tsai, "An efficient and accurate camera calibration technique for 3d machine vision," in *IEEE Conference on Computer Vision and Pattern Recognition*, Miami Beach, FL, 1986, pp. 364–374.

[10] S. Denman, V. Chandran, and S. Sridharan, "Robust multi-layer foreground segmenation for surviellance applications," in *IAPR Conference on Machine Vision Applications*, vol. 1, The University of Tokyo, Japan, 2007, pp. 496–499.